Analyzing NOAA Storm Damage Data

MATTHEW TUTTLE & LEX BUKOWSKI

Project Description

- NOAA Provides data on storms that have caused injuries or significant property damage dating back to 1950. The data set describes the location of event, type of weather, and magnitude of damage. There are approximately 1.8 MM weather events recorded.
- Interesting Questions
  - What type of weather is the most dangerous to people and property?
  - Is the magnitude property damage indicative of higher risk of injury to people?
  - Are certain locations prone to multiple type of dangerous weather events?
  - Does longer weather events lead to more property damage?
  - Have severe weather events increased over time?
- Prior Work:
  - "Student Research Abstract: Unsupervised Key Term Extraction of Tornado Narratives from NOAA Storm Events Database" –Emma Louise McDaniel
    - Narratives were text mined in order to retrieve the impacts of the disasters in order to structure the data for further use
  - "Some Comments on the Reliability of NOAA's Storm Events Database" –Renato P Dos Santos
    - Supported by limited statistical analysis, came to the conclusion that the database suffers from incompleteness and inconsistencies
- Datasets:
  - https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/
  - Data has been downloaded and combined on github project

Proposed Work

- Data Cleaning & Preprocessing
  - Data broken up into yearly reports which need to be combined
  - Ensure data types are consistent between files
  - Produce histograms and scatter plots to determine if there are any outliers and if any data fields are correlated
- Data Integration
  - There are more detailed data sets for location and fatalities that need to be tied together with the main data set.
  - There are codes for county and regional data that will need to be re-labeled for readability.
- Tools
  - We will use pandas and numpy to analyze the data. We will not be using a database to store the data as the size is reasonable to load directly.
- Evaluation
  - Are there “interesting” results produced by the data mining process